学校编码：10384　　分类号_____密级_____
学号：X2008221012　　UDC_____

厦门大学

工 程 硕 士 学 位 论 文

基于词频统计的自动文摘技术在音乐标签中的应用

Automatic summarization technology based on word frequency statistics and its application on music label

陈　昱

指导教师姓名：林琛　助理教授
专　业　名　称：计　算　机　技　术
论文提交日期：2011 年 10 月
论文答辩时间：2011 年 11 月
学位授予日期：2011 年　　月

答辩委员会主席：_____
评　　阅　　人：_____

2011 年 10 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（ ）1.经厦门大学保密委员会审查核定的保密学位论文，于

年　　月　　日解密，解密后适用上述授权。

（ ）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘 要

随着计算机科学和互联网技术的迅猛发展，网络中的各种多媒体信息近几十年来都在高速增长，对于音乐信息来说亦是如此。在面对数以亿计的音乐资源时，人们往往显得迷惘而不知所措。关于音乐的信息检索在近些年得到了计算机工作者的重视，基于内容，旋律等的音乐检索都获得了很大的进步。而在竞争日趋激烈的今天，各大网站纷纷完善自身的推荐系统，从以往的让用户找音乐，到如今的主动向用户推荐音乐。

本文的研究立足于完善音乐推荐系统，将基于词频统计的自动文摘技术应用于音乐标签的生成。通过对某张专辑的用户评论集进行文摘句的抓取，获得用户评论的精华部分，使之成为专辑音乐标签的内容之一。让用户更为快速，直观地获取评论信息。

在算法方面，本文先就词频统计的具体算法做详细的阐述。包括具体的流程、所需的数据结构等等，利用流程图和步骤分析详细说明算法的思路：按照先统计词频，再计算词权，最后计算句子权重的步骤完成自动文摘的生成。

在实现方面，本文结合不同的用户需求，设计出两种文摘生成方式。一种是基于单文档复合的方案。这是将所有用户的评论作为一个整体，生成文摘；而另一种是基于多文档的方案。这是将各个用户的评论单独作为一个生成文摘的单位，最后通过对各条用户评论的权重进行排序获得所需的结果。

最后，我们对两种方案的实验结果进行比较。并设计了下一步实验方案，以此作为本文的后续工作。与以往的研究不同的是，本文的创新点在于如何将自动文摘技术与音乐推荐相结合，并以此获得需要的结果。

关键词 音乐标签；自动文摘；词频统计

Abstract

With the rapid development of computer science and Internet technology, on the internet, multifarious information has increased rapidly in recent decades, it is also true for music resource. People often seem confused and overwhelmed when they face hundreds of millions of music resource. Computer workers are paying more and more attention to music retrieval, which is based on content, melodic, etc. and it has gained a lot of progress in recent years. With the fierce competition in nowadays, web sites are improving their recommendation system. Users find music resource in the past, and now, web sites take the initiative to recommend music to users.

This paper based on the improving music recommendation system, and applied the automatic summarization technology on generation of music label. Get the essence of user reviews, and make it as the contents of the music label by crawling digest in user reviews of an album. Make users get comment information more quickly and intuitively.

On algorithm, this paper first describes specific algorithm word frequency statistics in detail. Including specific process, data structures required, etc. Expound idea of the algorithm by analysis of the process and flowcharts: first, word frequency statistics, and then calculate the word's weight, finally calculate the weight of

the sentence to complete the automatic summarization.

On implementation, this paper designs two schemes with combination of different user needs. One is based on a single complex document. It sees all the user's comments as a group, generating digest; and another scheme is based on multi-document. It sees each user's comments as a separate unit to generate digest, and then get the desired result by the sort of the user comments' weight.

Finally, we compared with the results of the two schemes, and designed an experiment program as follow-up work for this paper. Different from previous studies, the innovation of this paper is how to combine the automatic summarization technology with music label, and get result.

Keywords Music label; Automatic summarization; Word frequency statistics

目 录

Contents

第一章 绪论

1.1 研究背景

信息技术的快速发展，尤其是网络技术和多媒体技术的发展，使得如今网络中的多媒体资源急剧增加。对于音乐爱好者来说，互联网所提供的海量资源和随心所欲的接收方式让音乐这一无国界的语言变得更加充满活力。成千上万的音乐网站在近十年中取得了飞速的发展，人们可以在线购买音乐 CD，可以在线收听音乐，也可以通过互联网将音乐下载到自己的随身听播放器里。同时，不受限于波段和信噪的影响，互联网用户可以通过互联网收听全球范围内的各大广播电台。互联网上海量的音乐资源给人们带了更加便利和随心所欲的娱乐生活，而随着信息技术的普及，音乐信息的传播量也呈指数级的增长，怎么在如此庞大的资源库中获得符合自己要求的信息是目前困扰许多互联网用户以及网络开发者的难题。

传统的音乐检索是基于文本信息的，也就是基于音乐的名称、歌手、词/曲作者或者歌词来进行标注，用户根据这些信息来进行检索，这样所获得的结果是比较精准的。另一种检索方式以地区、流派等比较大范围的属性作为检索关键字，这样检索得到的结果也比较海量，所以类似的检索通常带有一些盲目性，用户只是希望在一定范围内再进行筛选从而获得自己满意的检索结果。如今对音乐的检索已经非常人性化。但是，如何更好更精准地向用户推荐音乐依然是计算机开发者研究的课题。个性化的音乐网站在近些年有了长足的进步，对音乐信息的检索已经从以往单纯的分地区、分流派、分歌手到现在的分节奏、分情感、分群体等形式。

同时，网站为了获得更多的用户，应当尽可能减少用户自身的工作量，提供更为全面的推荐服务，让用户可以根据自己的喜好进行模糊检索，获得更多符合自身需求的推荐。这就需要给音乐加上标签，这些标签用来向用户展现音乐信息，从而让用户不需要听取整首音乐就能获得音乐的概况。

音乐标签作为音乐的“名片”，可以以多种形式存在。除文本之外，还有基于内容的音乐标签。而基于内容的标签主要是通过截取部分音频片段来向用户概

况并以此展示整曲的轮廓。目前国内外对此进行了大量的研究，总的概况起来可以分为两种形式：第一种是基于结构（Structure-based）技术，这种技术是将音乐中反复出现的片段作为该音乐的特征片段取出，作为标签。这样所截取的片段一般是一首音乐中的副歌部分。另一种技术是基于片段(Segmentation-based)技术，该技术先将整首音乐分割成若干片段，然后在每个片段中摘取具备典型特征的部分，从而形成一首音乐的预览。

音乐标签是网站向用户推荐音乐最直接地表达，如何用更加简短的信息更加全面精准地概括音乐是目前此类研究的直接目的。

1.2 研究的意义

与传统的文字资源不同，音乐资源是非结构化信息。目前网络音乐资源的描述和组织呈现出几个特点：一是资源类型多，资源量大；二是音乐描述的方案多而繁杂，没有统一的标准，给交流造成一定的不便；三是许多现存的描述方案并不能准确的揭示资源[1]。音乐标签的作用是让用户最快速地获取对一首未知歌曲或者一个未知专辑的了解，网站通过音乐标签向用户介绍音乐产品，以争取更多的用户。所以，对音乐标签的优化是非常有意义的：

（1） 自动生成的音乐标签可以大大提高效率。在世界范围内，每天都会有许许多多的音乐作品诞生，大型的音乐网站往往每天都要更新上线很多音乐作品，那么由系统自动生成准确的音乐标签就可以使网站的工作变得高效。

（2） 其次，清晰明了的音乐标签为用户节省了很多时间，从而提高用户选择的效率。用户在浏览音乐网站时通常要面对非常庞大的音乐资源，除了带有明确目的的检索以外，用户通常也会接受一些网站自身所作的推荐。而此时就面临着网站推荐信息的筛选问题。如何让用户快速地从海量的推荐中获取自己感兴趣的音乐是音乐网站所要解决的问题，音乐标签很大程度上是解决这个问题的方法之一。

（3） 最后，这本身也是与自动文摘技术相辅相成的。如今，电子出版物如潮而至，国际互联网蓬勃发展，大量的文献以机读的形式出现，网络上的信息

极大丰富。用户想要在信息的海洋中找到有用的信息，不仅需要先进的信息检索技术，还应当拥有一个能自动压缩甚至自动提炼信息的智能系统。而自动文摘技术正是符合这个背景要求的产物。利用自动文摘技术与音乐信息标签的相结合，我们可以同时促进两个研究方向的发展。

1.3 本文研究的主要内容和方法

1.3.1 课题的提出和主要内容

所有的音乐标签都是为了更准确地表达音乐信息。那么音乐标签应该提供哪些元素来概括音乐呢？歌曲名，作者，演唱者等标示性标签信息是非常具有客观性描述信息，而用户在欣赏某段音乐后给出的其他主观评价时常也能给别的用户带来有价值的参考。尤其是在某些具有代表性的音乐交流网站中，往往有非常多的用户进行评论和交流，那么这之中所产生的大量信息就十分具有价值。因此，我们将用户评论作为信息内容写入标签，以此作为音乐推荐的信息之一。本文以著名音乐网站豆瓣音乐（http://music.douban.com）作为研究的数据来源，进行文本音乐标签的研究。

要将非常分散的评论内容聚合到一小段概括性文字里，需要进行海量文字信息的精简和提炼。本文的研究就是如何从众多的用户评论中提取有价值的信息作为音乐标签。

1.3.2 研究的方法

将海量的用户评论精简成一小段具备典型特征的文字是本文研究的重点。本文的研究思路是将自动文摘技术结合到音乐标签的制作中。

自动文摘技术是信息时代发展到一定程度的必然产物，是计算机语言学和情报科学共同关注的课题，其本质是信息的挖掘和信息的浓缩。它能将人们从繁琐、冗余的信息中解脱出来，直接向用户提供简洁、信息全面的摘要，以提高用户获取信息的效率。

从理论上来说，对自动文摘的研究将有助于探讨人类理解、概括自然语言文

本，并从中获取知识的认识模型。自动文摘被认为是计算机实现自然语言理解的重要标志之一，因为它涉及到了大量的理论和应用技术，而且其相关理论方法和技术也可以应用到其他的自然语言处理应用领域，推动了整个自然语言处理领域的发展进程。从应用角度来说，在文献电子化和互联网迅速发展的今天，自动文摘系统的使用将大幅度降低编制摘要的成本，缩短文摘的出版周期，为人们廉价、迅速和准确地获得所需要的信息提供方便。

1.4 自动文摘技术概述和分类

文摘，是指全面准确地反映某一文献中心内容地简单连贯的短文。自动文摘技术用于自动从一篇或多篇文章中提取满足用户或应用需求的内容，加以组织后生成一篇内容完整、形式严谨的文摘[2]。自动文摘利用计算机自动地依据原始文献生成文摘，其中应包含原文的核心内容或者用户感兴趣的内容，以语义连贯的段落乃至篇章的形式输出，是用户快速获取感兴趣资源的一种准确高效的手段之一。

目前，自动文摘技术按照不同的划分方法可以有多种分类（见表 1-1）：

（1） 按文摘面向的用户划分，可以划分为通用型文摘和偏重型文摘。

（2） 按照文摘处理的文档数划分，可以划分为单文档文摘和多文档文摘[3]。

（3） 按照文摘生成所采用的方法分可以划分为摘录型文摘(Summarization Based On Extraction, SBE)、基于理解的文摘(Summarization Based On Understanding, SBU)、模板型文摘(Summarization Based On Template, SBT)和基于结构的文摘(Summarization Based On Discourse Structure, SBS)：

① 摘录型文摘(SBE)通常采用句子抽取方法生成文摘，该方法将文本视为句子的线性序列，将句子视为词的线性序列。主要步骤如下：首先计算词的权值和句子的权值，然后对原文中的所有句子按权值高低降序排列，权值最高的若干句子被确定为文摘句，最后将所有文摘句按它们在原文中的出现顺序输出。

② 基于理解的文摘(SBU)是以人工智能，特别是自然语言理解技术为基础而发展起来的文摘方法。这种方法与自动摘录的明显区别在于对知识的利用，它不仅利用语言学知识获取语言结构，更重要的是利用领域知识进行判断、推理得到文摘的意义表示，最后从意义表示中生成摘要。

③ 模版型文摘(SBT)有预先定义好的框架，文摘的生成过程其实就是从原文中检索出文摘模版所要求的内容，填到文摘模板中即可。其优点是比理解文摘中的脚本等要简单得多，更易于编写。缺点是，由于文摘框架的编写完全依赖于领域知识，所以信息抽取仍然受领域限制。文摘框架信息抽取要应用于多个领域，就必须为每个领域单独编写一个文摘框架，在处理文本时先进行主题识别，然后再根据主题调用相应的文摘框架。由于使用模板生成文摘，使得文摘的语言千篇一律，十分呆板。

④ 基于结构的文摘(SBS)采用自上而下分析方法，首先对文章的结构进行分析，再逐渐细化到段落、句子和概念，整个的分析过程是一个自上而下的过程，即由上层分析逐渐细化到底层分析。一般说来，文章中的不同部分承担着不同的功能，各部分之间在逻辑上是有一定的关联的。文章的这种关联找到了，其核心部分也就自然能够找到，这也就是基于结构的文摘的思想方法。应该说这种方法更利于从全局的观点把握原文作者的意图。但是目前来说，语言学对于文章结构的研究较少，可用的形式规则就更少了，这使得基于结构的自动文摘到目前为止还没有形成一套完整的理论方法。

表 1-1 自动文摘的分类

描述类别 / 划分标准	文摘类型	说明
按文摘面向的用户划分	通用型	针对一般用户而言，文摘反映的是文章的主体思想，作者的观点。
	偏重型	不区分单文档与多文档,根据用户需求为用户提供结果。提供的结果反映用户的兴趣和喜欢，而不是单纯反映作者的观点。
按处理的文档数目划分	单文档	针对单篇文档生成的摘要。
	多文档	将具有相同主题的多篇文档去除冗余、并生成一篇简明扼要的摘要输出。
按照文摘生成所采用的方法划分	摘录型	对原文的语句进行选择性的抽取获得摘要。
	基于理解	利用语言学知识获取语言结构，并且利用领域知识进行判断、推理得到文摘的意义表示，最后从意义表示中生成摘要。
	模板型	利用预先定义好的框架，从原文中检索出文摘模版所要求的内容，填到文摘模板中生成文摘。
	基于结构	分析文章的结构，利用各部分之间的逻辑关联找到核心部分，从而生成文摘。

1.5 自动文摘技术的国内外研究现状

国际上对自动文摘技术的研究可以追溯到上世纪 50 年代。1952 年,时为 IBM 公司研发工程师的 H.P.Luhn 开始研究通过计算机来为文本生成摘要的方法，经过 6 年多的研究，于 1958 年发表了其划时代的论文《The Automatic Creation of Literature Abstracts》，从此揭开了人们研究自动文摘的历史[4]。随后，马里兰州立大学的 Edmunds-on[5,6]、美国俄亥俄州立大学的 Rush[7]、英国兰开斯特大学的 Paice 等[8]选取字词的不同特征提取摘要。接着，开始有学者引入文档的结构特征和语义特征进行摘要的提取。美国耶鲁大学的 Schank[9]以及 GE 开发中心的 Rau 等[10]通过分析和推理得到文档的摘要。Sasha Blair-Goldensohm 等提出了 SC 算法[11]，该算法首先将句子进行聚类,根据每个类中的句子数目决定类的重要程度,再抽取重要的类的代表句子作为文摘。

从国外来看，对自动文摘的研究大体上有三个阶段：第一阶段是 1955 年至 1973 年的初始抽取时期；第二阶段是从 80 年代开始的人工智能方法时期，其中